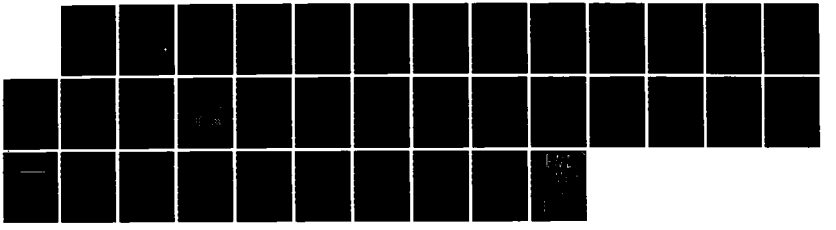
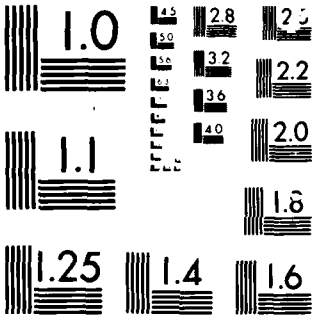
AD-A166 768 ARTIFICIAL INTELLIGENCE INDEXING: CREATING KNOWLEDGE BASES OF INDEX TERMS.. (U) HUGHES AIRCRAFT CO LONG BEACH CA SUPPORT SYSTEMS J J MYERS ET AL. DEC 85 1/1

UNCLASSIFIED HDL-CR-85-031-1 DAAK21-85-C-0031 F/G 5/2 NL





MICROCOPY CHART

HDL-CR-85-031-1

December 1985

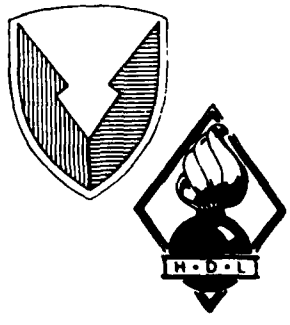
Artificial Intelligence Indexing:

Creating knowledge bases of index terms ordered by semantic relations

by Jay J. Myers, David Y. Kamemoto, and Arthur F. Griffin

Prepared by
Support Systems
Hughes Aircraft Company
Building A1, M/S 3C923
P.O. Box 9399
Long Beach, CA 90810-0399

Under contract
DAAK21-85-C-0031



U.S. Army Laboratory Command
Harry Diamond Laboratories
Adelphi, MD 20783-1197

Approved for public release; distribution unlimited.

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturers' or trade names does not constitute an official indorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

ADA166768

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) HDL-CR-85-031-1
6a. NAME OF PERFORMING ORGANIZATION Hughes Aircraft Company Support Systems 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION Harry Diamond Laboratories US Army Laboratory Command
6c. ADDRESS (City, State and ZIP Code) Bldg. A1 M/S 3C923 P.O. Box 9399 Long Beach, CA 90810-0399	7b. ADDRESS (City, State and ZIP Code) 2800 Powder Mill Road Adelphi, MD 20783-1197
8a. NAME OF FUNDING/SPONSORING ORGANIZATION US Army Materiel Command HDQ. 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER DAAK-21-85-C-0031
8c. ADDRESS (City, State and ZIP Code) 5001 Eisenhower Ave. Alexandria, VA 22333-0001	10. SOURCE OF FUNDING NOS. PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT NO.
11. TITLE (Include Security Classification) Artificial Intelligence Indexing (UNCL.)	
12. PERSONAL AUTHOR(S) Jay J. Myers, David Y. Kamemoto, and Arthur F. Griffin (Principal Investigator) HDL Contract John Carrier	
13a. TYPE OF REPORT Final Technical report 13b. TIME COVERED FROM 2/6/85 TO 12/31/85	14. DATE OF REPORT (Yr., Mo., Day) December 1985 15. PAGE COUNT 33
16. SUPPLEMENTARY NOTATION HDL Project No: AMS Code:	
17. COSATI CODES FIELD / GROUP / SUB. GR.	18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number) Information Retrieval, Indexing, Artificial Intelligence, Knowledge Representation, Natural Language Understanding

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This report summarizes an investigation of concepts related to the development of intelligent, automated text indexing. It discusses the construction of knowledge bases of index terms and semantic relations extracted from several different types of full-text documents. Approaches to automating this indexing process are examined, including work related to semantic knowledge representation, knowledge engineering, natural language understanding, and semantic inferencing. This research was conducted using a software package which uses semantic networks both for representing knowledge and for displaying a graphical index which can be browsed and manipulated to retrieve knowledge from an on-line documents data base. An approach to language processing is outlined which does not require a large lexicon but instead makes use of high-frequency, low-content words supplemented with knowledge of key verbs and generic and domain-specific seed knowledge. This approach uses word order, prepositions, and word endings to determine parts of speech, to identify index terms, and to infer the semantic relations among objects, modifiers, actions, and attributes.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS ☐	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE NUMBER (Include Area Code) 22c. OFFICE SYMBOL

DD FORM 1473, 83 APR EDITION OF 1 JAN 73 IS OBSOLETE.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

The parameters of the requisite seed knowledge and a framework for representing and making use of verb knowledge are described. This artificial intelligence approach to indexing is contrasted with human indexing performance.

CONTENTS

FIGURES

1. INTRODUCTION

Advances in modern computer technology have led to an exponential growth in the accumulation and storage of information. This is especially apparent in the scientific and technical literature, where both the supply and demand for information are rapidly expanding. DoD scientists, engineers, and technicians, as a result, are confronted with an overwhelming abundance of information. It is becoming increasingly difficult and time-consuming to locate and retrieve relevant information from the growing volume of technical literature and documentation. Similarly, it is becoming increasingly difficult and time-consuming to revise or update information in large text and graphics databases.

One might expect that the same technology which contributed to these problems would offer solutions. Indeed, a widespread transition is underway from paper-based documents to machine-readable textfiles, optical disks, and so forth, in order to allow on-line authoring, searching, and data modification. Yet current computer-based retrieval systems perform poorly and are difficult to use.[1] Furthermore, current systems deliver only bibliographic citations and abstracts, or at best the documents themselves, rather than retrieving information contained in the documents.

Since 1978, Hughes Support Systems has been engaged in an effort to resolve some of the problems associated with on-line technical documentation. In 1982 this effort began to focus on problems of information retrieval using artificial intelligence (AI) techniques to supplant inadequate conventional approaches. An IR&D project, known as Associative Loop Memory or ALOOP, was undertaken with the aim of developing a software/hardware system to extract and manage a database of index terms derived from natural language documents.[2] This retrieval system provides the user with a graphical display which can be browsed and manipulated to retrieve knowledge from an on-line documents database. An overview of this approach is depicted in figure 1.

---

[1] Landauer, T.K., Dumais, S.T., Gomez, L.M., and Furnas, G.W. "Human factors in data access." The Bell System Technical Journal, Vol. 61, No. 9, pp. 2487-2509, 1982.

[2] Griffin, A.F. "Intelligent information retrieval from on-line technical documentation." Proceedings of Joint Services Workshop on Maintenance Applications of Artificial Intelligence, Boulder, CO, Oct 4-6, 1983.

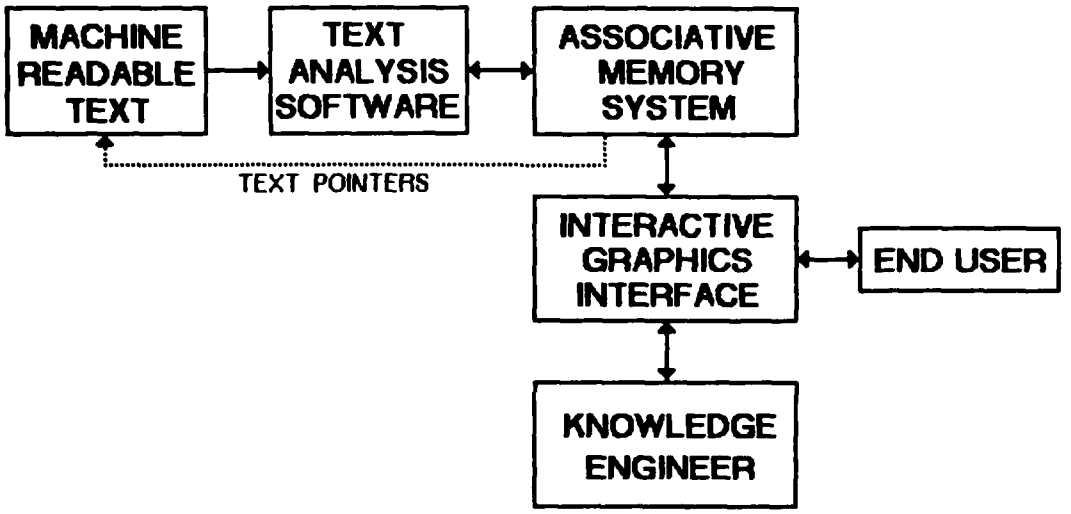


Figure 1. Overview of a system for automated indexing.

The term "knowledge engineering" refers to the procedures used by a trained user (the knowledge engineer) to manually insert knowledge, correct errors, or resolve ambiguities in the knowledge base. Recent efforts on the ALOOP project have been directed toward the production of software for natural language understanding and the addition of knowledge engineering tools. Although mutually reinforcing, the IR&D effort and the experiments performed under the present contract do not overlap.

1.1 Contract objectives

Under a previous contract, which was monitored by the Army Research Institute, we developed a knowledge representation scheme for indexing technical text and procedures for inferring semantic relations between index terms.

The aim of the present contract was to further investigate concepts related to the development of intelligent, automated text indexing. Our objectives were (1) to construct knowledge bases of index terms and semantic relations extracted from full text documents and (2) to investigate approaches to automating this extraction process including work involving semantic knowledge representation, knowledge engineering, natural language understanding, inferencing, and methods for handling uncertainty. These objectives were met through a number of experiments we conducted with small text databases.

1.2 Environment for conducting experiments

The experiments performed under this contract used the Interlisp-D language running on a Xerox 1100 AI workstation residing in our laboratory. Some of the experiments used proprietary software developed under the Hughes ALOOP project. The Xerox Lisp workstation is specially designed for AI developmental work and supports high-resolution bit-mapped graphics with windows and menus, uses a mouse for interacting with the screen, and has 29 megabytes of local disk storage. Our workstation is attached to a DEC VAX 11/780 through an Ethernet connection for remote file storage and access to textfiles, editors, and output devices.

1.3 Status of information retrieval

The traditional objective of technical libraries and of most other information retrieval efforts has been to deliver relevant documents or pointers to documents (citations) along with short summaries. The training of retrieval specialists

and the process of abstracting and cataloguing of information for these systems are difficult, expensive, and time-consuming. Moreover, the tasks of assessing the actual relevance of these documents and of finding and extracting the relevant information has been left to the user.

With the advent of inexpensive mass storage devices, it is becoming increasingly feasible to store the full text of documents in on-line databases. The development of devices for the rapid input of text, such as wordprocessors and optical scanners, has also contributed to the rising popularity of full-text databases. These advances make it considerably easier to physically store and retrieve text and graphics, yet the problem of identifying relevant knowledge in the database remains.[3]

Not uncommonly, the typical end user of an on-line database collection may have only a vague idea of what he or she is seeking and cannot provide a precise specification for conducting a search.[1] Even a well-defined query, however, must be translated into a more constrained and often more ambiguous format for input to the system, usually by a trained human intermediary. Indeed, a significant area of research in information retrieval is concerned with automating the query translation and narrowing operations.[4,5] Another similar

---

[1]Landauer, T.K., Dumais, S.T., Gomez, L.M., and Furnas, G.W. "Human factors in data access." The Bell System Technical Journal, Vol. 61, No. 9, pp. 2487-2509, 1982.

[3]Blair, D.C., and Maron, M.E. "An evaluation of retrieval effectiveness for a full-text document-retrieval system." Communications of the ACM, Vol. 28, No. 3, pp. 289-299, 1985.

[4]Salton, G., Buckley, C., and Fox, E.A. "Automatic query formulations in information retrieval." Journal of the American Society for Information Science, Vol. 34, No. 4, pp. 262-280, 1983.

[5]Tou, F.N., Williams, M.D., Fikes, R.E., Henderson, A., and Malone, T.W. "RABBIT: An intelligent database assistant." Proceedings of the National Conference on Artificial Intelligence. Pittsburg, PA. August 1982.

approach constructs relational thesauri to improve retrieval performance in response to user queries.[6]

Current on-line information retrieval systems illustrate these problems. One common approach is to present to the user a nested menu display which contains subject headings similar to those found in a traditional subject index of a printed document. However, not only is this type of index expensive and time-consuming to produce, but it is again of little use to a user who does not know or cannot find the subject heading for the information she or he is seeking.

Another common retrieval approach uses keyword query techniques. These systems require sophisticated search strategies involving the selection of appropriate index terms and Boolean operators. As a result, most searches are conducted by highly trained intermediaries rather than by untrained end users.[7] Furthermore, it is often difficult to narrow a search with these methods; a given query may retrieve either hundreds of documents or none. Query systems focus on only two of the many possible retrieval cues found in full text documents--the frequency and proximity of word occurrences. Therefore, the response to a word like "terminal" may include instances ranging from "computer terminal" to "terminal disease" to "airline terminal" constrained only by the breadth of the database (and by subsequent Boolean quantifiers).

2. AN AI APPROACH TO INFORMATION RETRIEVAL

We can perhaps best define our AI-based approach to the information retrieval problem by contrasting it with the traditional systems described above. We intend to develop a system that does not require exact knowledge of the objective of a search but can be used by an unsophisticated user to rapidly retrieve relevant sections from a large on-line collection of full text documents. We assume that the user will recognize the information when retrieved. The system

---

[6] Wang, Y., Vandendorpe, J., and Evens, M. "Relational thesauri in information retrieval." Journal of the American Society for Information Science. Vol. 36, No. 1, pp. 15-27, 1985.

[7] Eisenberg, M. The direct use of online bibliographic information systems by untrained end users: A review of research. Information Resources Publications, Syracuse, N.Y., 1983.

will allow the user to quickly judge relevance by delivering specific passages of the text rather than entire documents. Alternative search paths will be provided if a search fails to retrieve the desired information. Further, the system will be dynamic; over time, it will improve its ability to process text and to retrieve relevant information by making use of knowledge supplied by a knowledge engineer, information from text it has previously read, and interactions with end users. Finally, the system will extract index terms and semantic relations from the text automatically in order to overcome the bottleneck of knowledge acquisition which plagues most current AI systems. In brief, our objective is to present the user with an easily understood graphics interface which can be used to intelligently query and browse through a massive text database.

2.1 Semantic network as a browsing index

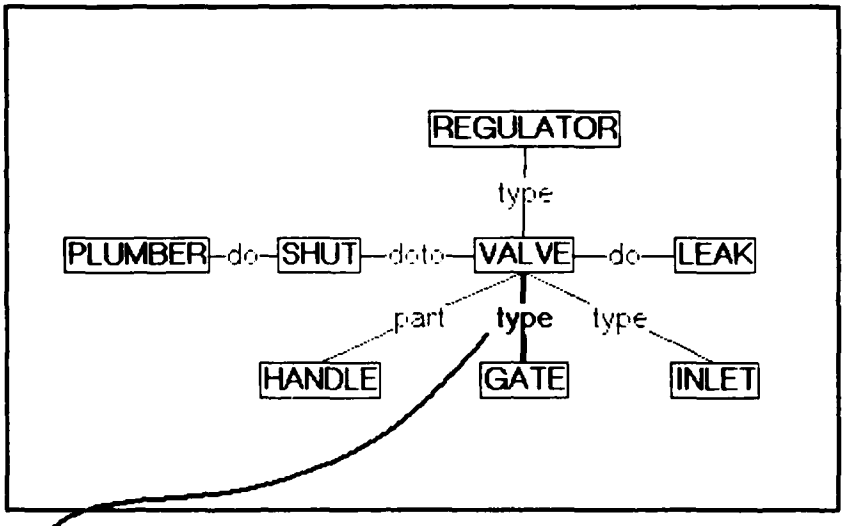
Semantic networks are a common technique for representing knowledge in AI systems. Originally proposed as a model of human associative memory,[8] they have come to be regarded as convenient structures for representing certain kinds of knowledge and for organizing this knowledge in a way that is convenient for making inferences. Semantic networks represent knowledge as concepts linked together by semantic associations.[9] Associative memory recall can then be modeled as the traversal of these links.

Figure 2 depicts a small portion of a sample semantic network used for browsing text. Notice that this network is structured so that vertical links represent a hierarchy of information from general to specific (e.g., VALVE type REGULATOR) while horizontal links represent agent-action-object (PLUMBER do SHUT doto VALVE) or object-attribute-value relations. For the net to be used as a text index, we attach pointers to the semantic links corresponding to sections of text that contain the information underlying the semantic relations represented in the network. The user, by selecting two linked index terms with a mouse or

---

[8] Quillian, M.R. "Semantic memory." In M. Minsky (Ed.), Semantic Information Processing. MIT Press, Cambridge, Mass., 1968.

[9] Brachman, R.J. "On the epistemological status of semantic networks." In N.V. Findler (Ed.), Associative Networks--Representation and Use of Knowledge by Computers. Academic Press, N.Y., 1979.



When water breaks loose, the first thing to do is shut it off at the source. If you have time to think, pick the valve nearest the leak. **But if things are moving too fast, head first for the gate valve and shut down the whole system while you plot the rest of your attack.**

These are typical valves. Know where to find the ones in your house.

Figure 2. Semantic network as a browsing index; selecting a link in the semantic network retrieves a corresponding full text passage.

other input device, can retrieve the text passage corresponding to the pointer. This approach results in a very browsable index ordered by the natural relations beween index terms, rather than by the arbitrary alphabetical ordering found in standard printed indices.

2.2 Semantic inferencing

While a traditional index is a very simple representation of the knowledge in a text, a semantic network representation is much more structured and useful for performing inferences. A semantic network typically employs a set of semantic primitives which specify the types of relations which can be represented in the system. Our system uses a set of 16 primitives (type, inst[ance], part, do, doto, has, is, and mod[ifies] and their negations) to capture relationships involving class membership, part/whole relations, actions, and properties. Each node taken in conjunction with its relations can be thought of as a concept or a frame (another common form of AI knowledge representation).

The vertical membership relations in a semantic network permit the system to make logical inferences using the knowledge it has stored. For example, if a network represents the relations that a robin is a type of bird and that all birds have feathers and can fly, then the system can infer that a robin has feathers and can fly. Similarly, from the relations that a plumber can repair any plumbing device and a faucet is a plumbing device, the system can infer that a plumber can repair a faucet. This inferencing ability permits a more efficient knowledge representation, in which attributes common to all members of a class can be attached to the class rather than being duplicated for each member. Further, the system can use inferences during the processing of text to understand new words and to relate knowledge from the text to knowledge already stored in the system.[10]

2.3 ALOOP hardware

The semantic networks constructed to index and represent the knowledge in a large document collection will obviously need to be very large. Real-world databases typically contain thousands of documents pertaining to a large variety of

---

[10] Kiersey, D.M. Word learning with hierarchy-guided inference. Ph.D. Dissertation, Department of Computer Science, University of California, Irvine, 1983.

subjects. Conventional serial computers may not be able to perform searches of these networks rapidly enough to provide a reasonable response time.[11] To cope with this problem, we are also exploring, as part of our IR&D effort, the implementation of the ALOOP concept as special-purpose, parallel-processing hardware.

The large semantic network would be segmented and stored in a number of independent processing nodes, each of which would contain a microprocessor and local memory (fig. 3). Searches could then be performed in parallel, simultaneously in each of these nodes, rather than by serial searching of the entire network. A small working hardware prototype was developed in 1982, and we anticipate a further hardware effort as we begin to work with larger databases.

3. CREATION OF KNOWLEDGE BASES

One of our objectives under the present contract was to develop and study knowledge bases of index terms and semantic relations extracted from different text domains. Three general domains were chosen for analysis: research abstracts, maintenance manuals, and encyclopedic texts. Knowledge bases were created for these domains using representative documents--for the abstracts collection, DoD form 1498s pertaining to training; for the maintenance domain, a depot maintenance work requirement of a TOW weapons system and sections of a home maintenence manual; and for the general text, a short segment from a geology textbook. This latter text was chosen to allow comparison of our approach with a published study of human indexing performance[12] (sect. 4.1 below).

The process of creating knowledge bases was very different for these domains. Creating a knowledge base from the abstracts collection proved to be most difficult. This difficulty apparently stems in part from the fact that abstracts are abbreviated texts rather than full text documents. The language used in abstracts is condensed and

---

[11] Fahlman, S.E. "Representing and using real-world knowledge." In P. Winston & R.H. Brown (Eds.), Artificial Intelligence: An MIT Perspective, pp. 454-470. MIT Press, Cambridge, Mass, 1979.

[12] Jones, K.P. "How do we index: A report of some ASLIB informatics group activity." The Journal of Documentation, Vol. 39, No. 1, pp. 1-23, 1983.

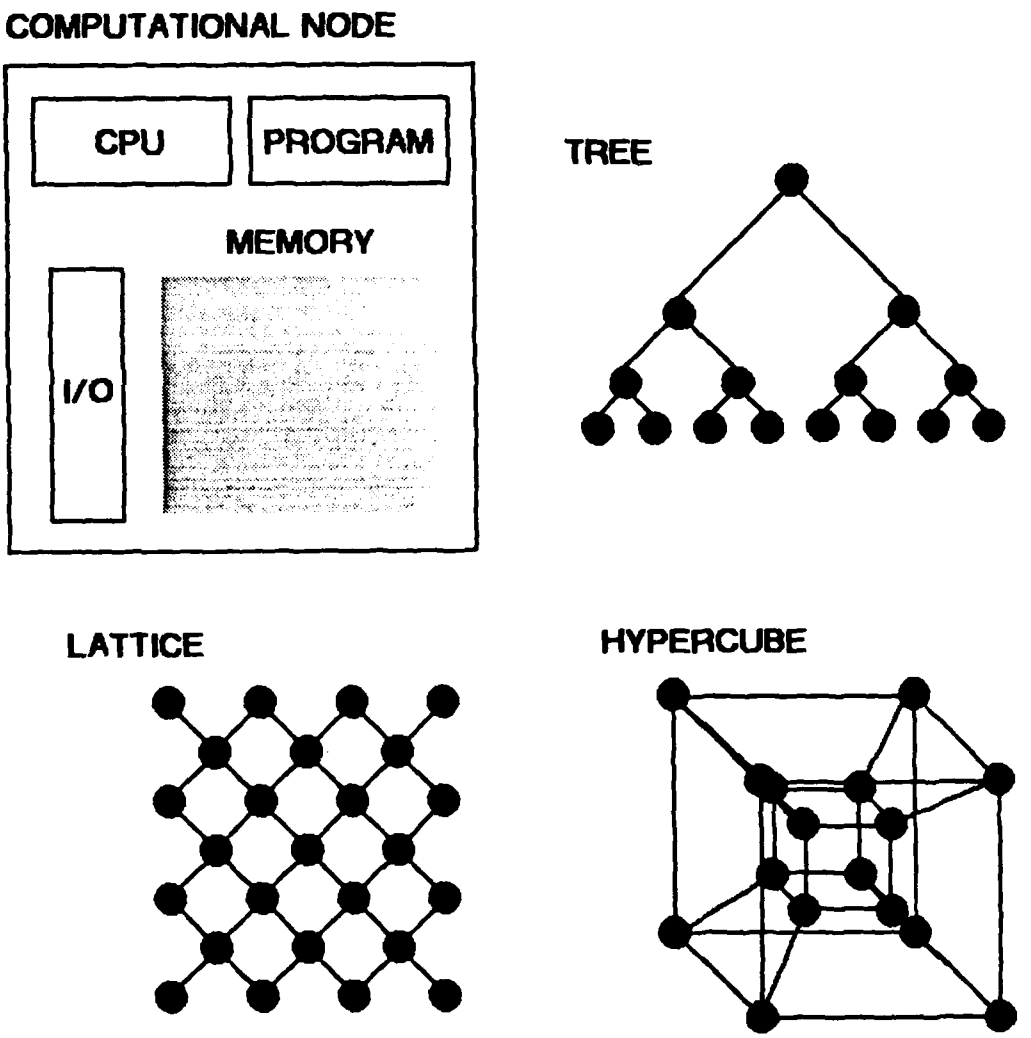


Figure 3. Associative Loop Memory parallel-processing node and message-passing architectures.

unnatural; there is a tendency to use longer words, more complicated noun phrases (e.g., an engagement simulation-based army training evaluation program) and long, convoluted, and incomplete sentences. Even humans often have difficulty understanding these abstracts. Some examples are presented in figure 4.

Even more problematic from the standpoint of indexing is the absence of terms which can be used to identify the content of the abstracts. The research summaries tend to be written in a very general and abstract language. Yet another problem with this domain was the use of a standardized format which often caused the abstracts to read alike.

The language found in maintenance manuals is also constrained and unlike that generally encountered in text. By necessity the information contained in these documents is highly process-oriented and presented in terms of symptoms and corrective actions or as specific sequential instructions. These documents also tend to use domain-specific terminology and to emphasize spatial relations (i.e., where to find a part in relation to other parts) which are not represented in our semantic primitives. Another problem is the reliance upon tables and figures to convey information, which cannot be captured by straightforward text analysis procedures.

The general expository text was most readily processed by our approach. However, the effort with this text also highlighted the need to incorporate very low frequency terms, to handle modifier-noun phrases, and to infer relations involving terms which may not be found in the text. More discussion of this text is presented in section 4.1.

In general, we concluded from our experience with these texts that initial generic and domain-specific knowledge was needed to accurately process and represent the texts. Appropriate general knowledge of of objects, actions, and attributes, as well as knowledge of the important concepts and relations of a specific domain can be provided by a knowledge engineer to greatly facilitate the analysis of a text for index terms and semantic relations. We refer to this initial background knowledge as "seed knowledge." The construction of hierarchically ordered networks for browsing through the database is critically dependent upon the scope and accuracy of the seed knowledge. We also estimated the optimal length of a text sample for indexing to be on the order of 2,000 to 10,000 words.

Create a methodology for the development of "relevance values" to career enhancement and for the definition of functional overlap of separate external-to-specialty assignment options for the career progression demands of the individual officer in his own assigned specialties; develop an assignment algorithm methodology.

Evaluate the potential of quasi-algorithm methods and techniques for specifying objective job/task descriptions and performance requirements that are related to job structures, work requirements, training, and personnel management.

Systematically investigate organizational effectiveness (OE) as a process, including principals within it and the nature of their behavioral dynamics; develop an operational description of the conditions for and dynamics of the OE process.

Improve individual and unit proficiency of army personnel in selected military systems by developing guidelines and recommendations for the effective transfer of training technology to the army user.

Develop a productivity measurement system for use in civilian personnel management to provide input into training for civilian and military managers on the relationship between personnel management practices and mission accomplishment.

Figure 4. Excerpts from abstracts database of form 1498s.

3.1 Seed knowledge experiments

From our experiments, we can begin to characterize what information the seed knowledge should contain. Most important for organizing the knowledge base is an initial set of very abstract terms pertaining to the domain or to general concepts. These terms may rarely be encountered in text but they provide a taxonomy of objects which is useful for browsing and for making inferences. Ideally, this taxonomy should be carried down to the level of the major keywords which do occur in the text. An example for the home plumbing repair text is depicted in figure 5. A knowledge engineer does not necessarily need to provide extensive seed knowledge before a text is analyzed, but can be examine the semantic networks created after the text has been processed and iteratively built up this knowledge. Also, once this hierarchy of seed knowledge has been built up for a given domain, it should be applicable to additional documents from that domain. The text analysis procedures should thus become increasingly effective and require less intervention.

3.2 Identification of index terms

When a text is viewed simply as symbol strings, a number of cues become apparent which might help to identify the key terms and their meaning. These include:

document cues--headings and paragraph breaks
sentences breaks and punctuation
sentence length
word frequency
word length
word position within a sentence, paragraph or document
word order and interproximities--phrase groupings
phrase frequency and length
word morphology--prefixes and suffixes

Higher level semantic cues include:

word function--syntax
word meaning--semantics
context or theme

The above cues are roughly ordered from rather weak, global cues to those which are stronger and more specific. The former are more concrete and relatively easy to compute, whereas the latter are abstract, computationally difficult or costly, often requiring external knowledge.

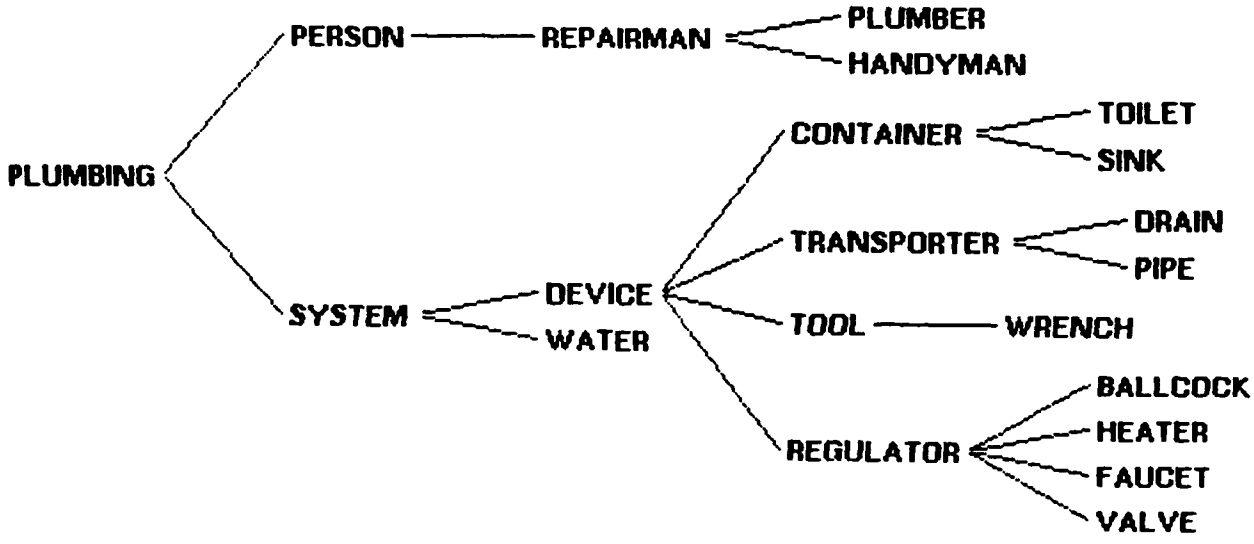


Figure 5. Hierarchy of seed knowledge for the plumbing domain.

For the purpose of representing the key knowledge in a text and creating an index of key words and semantic relations, we do not need to have an in-depth understanding of the text. Rather, it is sufficient to obtain a limited understanding which focuses on key words and concepts. Important concepts in terms of indexing are likely to appear in a context which is easy to understand and to occur with some frequency so that difficult passages can be ignored. The following is an overview of the approach we used in creating the knowledge bases.

3.3 Word frequency and semantics

From an examination of the most frequent words found in studies of English text,[13] it can be seen that words that occur with high frequency generally do not convey much meaning (although they are important syntactic markers). Figure 6, for example, depicts the 50 most frequent English words from the Kucera-Francis study. For purposes of indexing, these words often are and should be ignored. This simple fact is very important because it greatly reduces the problem of indexing. As an illustration, the most frequent 100 English words account for nearly half of all of the words encountered in a typical text. Furthermore, it is quite easy to automate this task of omitting highly frequent words from an index by use of an exclusion list ("stoplist").

A correlary of the inverse relationship between frequency and semantics is that words which occur very infrequently may be critical for understanding the meaning of a sentence. Thus, although there are a limited number of articles, conjunctions, or prepositions in English, there is a profusion of nouns and verbs which are used to express subtle shades of meaning or domain-specific terminology. The manual encoding of the semantics of these words is costly and laborious and this is compounded by their large number and relatively low frequency. These are the primary terms we seek to identify and use for indexing.

3.4 Frequency experiments

A word-frequency analysis of the databases confirmed the validity of the use of a stoplist of frequent terms with low semantic content. For example, the most frequent words in

---

[13] Kucera, H. and Francis, N. Computational Analysis of Present-Day American English. Brown University Press, 1967.

1.	the	26.	from
2.	of	27.	or
3.	and	28.	have
4.	to	29.	an
5.	a	30.	they
6.	in	31.	which
7.	that	32.	one
8.	is	33.	you
9.	was	34.	were
10.	he	35.	her
11.	for	36.	all
12.	it	37.	she
13.	with	38.	there
14.	as	39.	would
15.	his	40.	their
16.	on	41.	we
17.	be	42.	him
18.	at	43.	been
19.	by	44.	has
20.	I	45.	when
21.	this	46.	who
22.	had	47.	will
23.	not	48.	more
24.	are	49.	no
25.	but	50.	if

Figure 6. Most frequently occurring English words in printed text.
[from Kucera, H. and Francis, N. Computational Analysis of Present-Day American English. Brown University Press, 1967.]

each of the domains showed a close correspondence with the Kucera-Francis list. Moreover, many high frequency words which were not on this list were indicative of the domain (e.g., research, training, plumbing, valve, lava, etc.). We can conclude that the stoplist significantly reduces the indexing task and that the remaining frequent words are often key concepts in a domain. This is an important point for attempts to provide seed knowledge for a domain. However, given that frequently occurring words will have more opportunities to be extracted during the processing of a text, there may be no need to explicity monitor word frequency.

The frequency analyses also revealed that nouns occurred with a much greater frequency that verbs. Moreover, each domain tended to use a given subset of verbs which thus provide another key to understanding the text within that domain.

In an additional frequency study, we attempted to identify common noun phrases (e.g., utility company serviceman) by counting co-occurrances of nonstoplist words. Many of the important indexable phrases were identified by this method. However, the computational expense of maintaining these frequency counts was too high. It appears that some other way of incorporating these phrases as index terms is needed.

3.5 Word stemming

Keyword indexing systems typically remove the suffixes of words, thus grouping together similar word forms.[14] This allows a more compact index and permits a single query to retrieve all the various word forms. A problem arises, however, in that stemmed words are more ambiguous (e.g., calculat-ion and calculat-or become synonymous). Our experience with the frequency experiments and knowledge base creation process suggested that some stemming was necessary to avoid duplication of index terms. However, we concluded that stemming could be limited to conversion of plural nouns to their singular forms and to conversion of marked verb forms (verbs ending in "ing," "ed," or "s," and those formed irregularly) to unmarked forms.

---

[14] Dillon, M. and Gray, A.S. "FASIT: A fully automatic syntactically based indexing system." Journal of the American Society for Information Science, Vol. 34, No. 2, pp. 99-108, 1983.

4. TEXT ANALYSIS

Most language understanding efforts use a large dictionary or lexicon (on the order of 30,000 words or more). Yet these systems often encounter words which are not represented in the lexicon, especially within domains which use special terminology. No fixed set of words and word usages can ever completely encompass a natural language because of the dynamic and flexible nature of natural languages.

Our approach to text analysis is robust because it avoids the use of such a large lexicon. We attempt to provide an exhaustive knowledge of articles, conjunctions, copulas, functionals, prepositions, and pronouns, largely corresponding to the most frequent English words. The identification of adjectives, adverbs, nouns, and verbs, however, is left to the language processor. This, of course, requires the ability to appropriately process new words for which no prior knowledge has been stored. However, such a system can build up a lexicon as it reads text and so naturally customizes the lexicon to the text domain.

Our language processor uses word order, prepositions, knowledge of selected verbs, and word endings to identify the parts of speech and to infer the semantic relations among modifiers, objects, actions, and attributes. For example, it infers that any word which immediately follows an article must be an adjective or a noun, that a word ending in "ly" is most likely an adverb, that a noun followed by a verb should be represented by a do relation, and so forth. In addition, certain key phrases which directly express important semantic relations (e.g., "a ‹noun phrase› is a type of ‹noun phrase›" or "a ‹noun phrase› is composed of ‹noun phrase›") are specially processed.

4.1 Comparison with traditional indexing

In order to evaluate the promise of an AI approach to indexing, we wanted to contrast our approach with expert human indexing performance. However, little has been published regarding the performance of human indexers. We could find no standards for evaluating current indexes, few published research reports, and only a few standard guidelines used by trained indexers.

One research report, however, attempted to contrast the performance of a number of trained indexers working with the

same short segment of text.[12] The sample text from this experiment is depicted in figure 7 along with sections of the actual book index which referred to this text. The major conclusions of the study can be briefly summarized here. Seventeen human indexers selected a total of 37 different index terms. The indexers chose 16 different phrases (e.g., silica rich lavas) as index terms. These included terms which occurred very infrequently or not at all in the text (volcano). The frequency of the selection of a few key terms by many of the indexers did indicate some agreement. However, for the most part, these results, like the book index, showed human indexing to be highly idiosyncratic.

We attempted to develop a semantic net index for this same text by semiautomatically applying the text analysis techniques described above. The resulting display is shown in figure 8. It can be readily seen that our approach requires much less information to capture most of the references incorporated in the indexes created by hand. Furthermore, the semantic network is a more organized representation of the knowledge contained in the text. A reference to volcano was included as a seed relation. We are encouraged by this validation of our approach.

4.2 Verb knowledge

Case grammars parse a sentence on the basis of functional roles rather than parts of speech.[15] These grammars incorporate knowledge about key verbs to restrict the candidates words which can fill the roles of agent, object, instrument, and so forth for a given verb. These cases may be considered equivalent to the slots of a frame representation.[16]

---

[12] Jones, K.P. "How do we index: A report of some ASLIB informatics group activity." The Journal of Documentation, Vol. 39, No. 1, pp. 1-23, 1983.

[15] Fillmore, C. "The case for case." In E. Bach and R. Harms (Eds.), Universals in Linguistic Theory, pp. 1-88. Holt, Rinehart and Winston, New York, 1968.

[16] Charniak, E. "The case-slot identity theory." Cognitive Science, Vol. 5, No. 3, pp. 285-292, 1981.

LAVAS

from A. Holmes:
Principles of Physical Geology

The temperature of freshly erupted lava is rarely much above the melting-point, and according to the composition and gas content it may range from 600 to 1200@C, the basic lavas, like basalt, being generally the hottest. The mobility of molten lava depends on the same factors. Silica-rich lavas are usually stiff and viscous and congeal as thick tongues before they have travelled far, whereas basic lavas tend to flow freely for long distances, even down gentle slopes, before they come to rest. The speed of a lava stream depends on the mobility and slope, and may, quite locally, reach 50 miles an hour. But such speeds are very rarely attained; even 10 miles an hour is unusual and often the movement is sluggish. In recent years, when approaching flows have threatened villages on the slopes of Mauna Loa and Etna, the danger has been averted by bombing from aeroplanes, whereby the flows have been constrained to follow new and less menacing courses.

The surfaces of newly consolidated lava flows are commonly of two contrasted types, described in English as block and ropy lavas, but known technically by their Hawaiian names, aa (ah-ah) and pahoehoe respectively. Block lava forms over partly crystallized flows from which the gases escape in sudden bursts. During the advance the congealing crust breaks into a wild assemblage of rough, jagged, scoriaceous blocks. Ropy lava begins at a higher temperature; minute bubbles of gas escape tranquilly and the flow congeals with a smooth skin which wrinkles into ropy and corded forms like those assumed by flowing pitch. It sometimes happens that after the upper surface and edges of a flow of this kind have solidified, the last of the molten lava drains away, leaving an empty tunnel. Some of the lava caves of Iceland are famous fro the shining black icicles of glass which adorn their roofs.

When lava of the ropy type flows over the sea floor, or otherwise beneath a chilling cover of water, it consolidates with a structure like that of a jumbled heap of pillows, and is then appropriately described as pillow lava. By the time each emerging tongue of lava has swollen to about the size of a pillow the rapidly congealed skin prevents further growth. New tongues which then exude through cracks in the glassy crust similarly swell into pillows and so the process continues. The structure is a common one in the submarine lavas associated with the geosynclinal sediments of former periods, and has been seen actively developing in modern flows that reached the sea floor. Columnar structure develops within the interior of thick masses of lava which have come to rest and have consolidated under stagnent conditions. It is especially characteristic of very fine grained plateau basalts which are relatively free from vesicles.

Figure 7. Traditional indexing example.
[from Jones, K.P. "How do we index: A report of some ASLIB informatics group activity." The Journal of Documentation, Vol.39, No.1, pp. 1-23, 1983.]

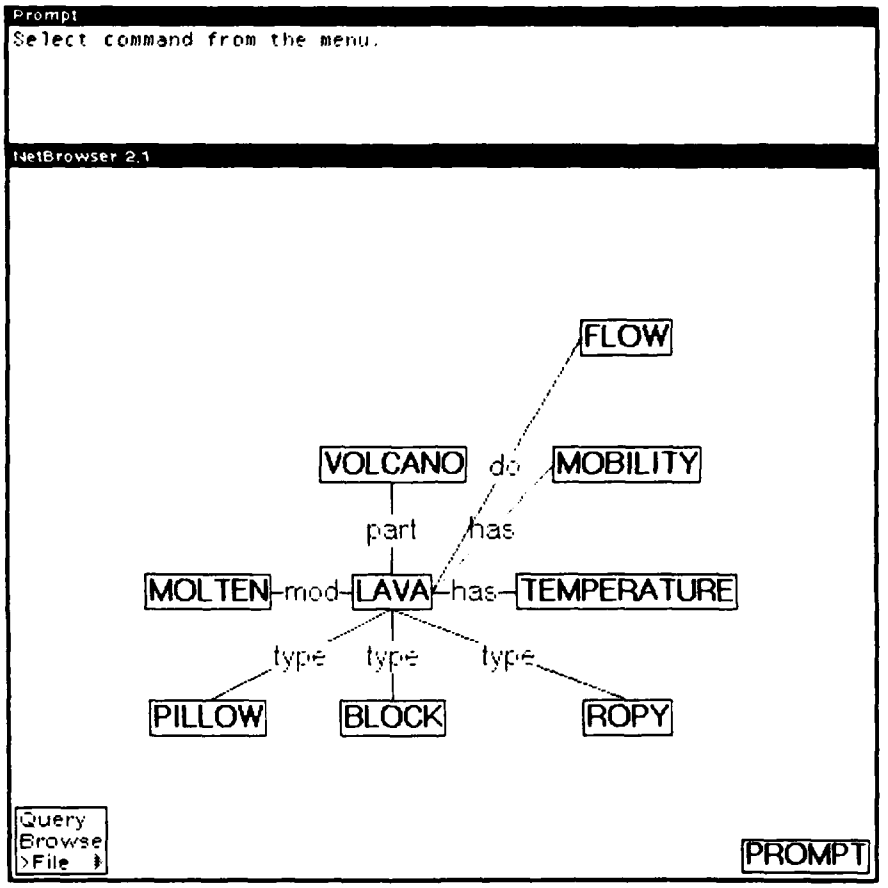


Figure 8. Semantic network indexing example.

Schank has extended the idea in his conceptual dependency theory[17] to permit inferences, which pertain to the motives of the participants and the implicit details of what has transpired, to be drawn from common actions. He represents the events underlying verbs as action primitives with a specified actor and object and a direction of action. For example, the transfer of possession is one such action primitive, which Schank calls ATRANS, used to encompass such verbs as give, take, receive, and sell. By recognizing a verb to be a type of ATRANS, the meaning of the sentence can be unravelled.

Under this contract, we have explored the application of similar knowledge about verbs to the processing of text. While it is not our objective to achieve an in-depth understanding, our experiments suggest that it is important that seed knowledge of the key verbs in a text domain be provided. Verb knowledge is central to the task of language understanding at even a shallow level and permits the formation of inferences which can link terms to the seed knowledge object hierarchy.

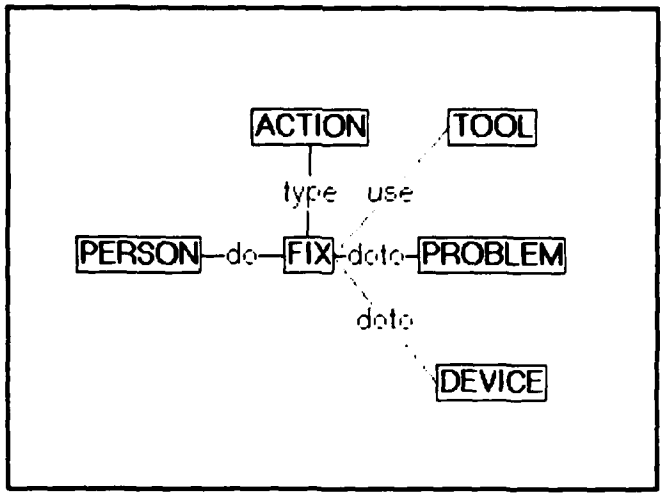
In our system, verb knowledge consists of the root verb forms along with the likely categories of actors and objects. As the example in figure 9 shows, this verb knowledge is represented as a semantic network. When a form of the verb is encountered in text, the system can then infer that the actor and object are members of the specified categories. In the example, from knowledge of the verb fix, the actor is inferred to be a person and the object to be a device or problem. The network representation permits creation of a verb hierarchy with inheritance of the inferential knowledge. Thus, for example, to add knowledge of the verb repair to the system, it would be necessary only to specify that it is a type of fix.

4.3 Disambiguation of multiple word meanings

The problem of multiple word senses or ambiguity pervades all efforts to understand text. A simple word like "type" can be used in a variety of ways which can be differentiated and understood only on the basis of context cues or other knowledge. Conservative estimates suggest that over 30 percent of occurrances of English words are lexically

---

[17] Schank, R.C. Conceptual Information Processing. North Holland Publishing, Amsterdam, 1975.



The foobaz fixed the framis with a widget.

(FOOBAZ type PERSON)

(FRAMIS type PROBLEM) or (FRAMIS type DEVICE)

(WIDGET type TOOL)

The accountant fixed his mistake.

John is fixing the leak.

An electrician could fix the switch with a screwdriver.

Figure 9. Representation of verb knowledge with examples of application.

ambiguous.[18] This problem is thus not resolved by reliance upon a large lexicon of words. Rather, the solution requires a greater semantic understanding of the text. This can be achieved by constraining the word meanings with more extensive domain knowledge, by mechanisms to track and make use of context,[19,20] by logical validation techniques, or by intervention of a knowledge engineer. We have begun to conduct preliminary experiments is this area.

4.4 Computerized writing aids

The clarity of technical writing, whether of research reports, documentation or abstracts of research, is generally more of a problem than the content. Current efforts to improve the clarity of technical writing tend to be ineffective because they rely upon editing or revision by someone other than the original author or because they use guidelines which are difficult to apply or which are based on faulty or imprecise assumptions. Present computerized writing aids provide an author with limited feedback largely restricted to statistical analyses or to analyses of individual sentences in isolation. For example, these systems can detect words not in a standard vocabulary, compute sentence lengths, flag use of the passive voice and so forth.

A new approach to enhancing the clarity of technical writing, which relies upon AI techniques for understanding natural language, has been recently outlined by Kieras.[21] This approach uses a computerized system to scan a text and detect problems in the writing. Correction of the text is left to the author. The rules used by such a computerized system,

---

[18] Britton, B.K. "Lexical ambiguity of words used in English text." Behavior Research Methods & Instrumentation, Vol. 10, No. 1, pp. 1-7, 1978.

[19] Alshawi, H. A mechanism for the accumulation and application of context in text processing. Technical Report No. 48., University of Cambridge Computer Laboratory, 1983.

[20] Charniak, E. "Passing markers: A theory of contextual influence in language comprehension." Cognitive Science, Vol. 7, pp. 171-190, 1983.

[21] Kieras, D.E. "The potential for advanced computerized aids for comprehensible writing of technical documents." Technical Report No. 17 (TR-85/ONR-17), Office of Naval Research, 1985. (ADA150501)

however, would be based on findings from research involving the psychology of comprehension. Kieras contends that these findings provide a more suitable set of guidelines for improving the clarity of writing than the guidelines which are currently used.

This new approach to developing advanced computerized writing aids interests us for several reasons. It utilizes a semantic network representation of the text to track and detect difficulties in the writing. It is also consistent with our contention that in-depth text understanding is beyond the current state-of-the-art of natural language understanding systems. Like our work on indexing, such a computerized writing aid would not require an in-depth understanding of the text because the goal is simply to identify when the comprehension of the text is difficult. Futhermore, much of our effort to develop language understanding algorithms for the identification of index terms and semantic relations could be carried over to the development of algorithms for detecting and improving the clarity of writing.

4.5 Knowledge-based input assistant

Through our experiments with the form 1498 database and conversations with the Defense Technical Information Center, we have identified some overall shortcomings of the 1498 abstracts (some of which were mentioned above) which contribute to difficulties in indexing and retrieving these research summaries. We have noted an important connection between the quality of data input to a database and the ability to subsequently organize and retrieve this information. The input problems for the 1498s fall into several categories: (1) a lack of adequate guidelines or feedback for effectively completing the form, (2) a failure on the part of the researcher to perceive the utility of conscientious completion of the form, (3) inconsistant or poor usage of language, and (4) inadequate knowledge of related research.

We believe the quality of the input could be improved by an on-line knowledge-based assistant which would automatically identify shortcomings during the input process and provide knowledge, feedback, or further processing of the text to rectify these problems. This assistant would be linked to our information retrieval system and would have, in addition, knowledge of the 1498 format, of language syntax and semantics, and of the specific research domains being input. Knowledge from other abstracts would be retrieved for use in guiding the input and to inform and hopefully motivate the author.

5. CONCLUSIONS AND RECOMMENDATIONS

We have carefully studied the characteristics of English documents of several different types to determine the most important problems confronting an effort to automatically index text by extracting keywords and semantic relations. The process of creating such knowledge bases was found to vary according to the particular text domain. We found that this effort requires an initial kernal of seed knowledge, particularly knowledge of the specific domain.

Our experiments suggested an approach which makes use of high-frequency, low-content words supplemented with knowledge of key verbs and seed knowledge. Under this contract we developed these concepts by defining the parameters of the requisite seed knowledge and creating a framework for representing and making use of verb knowledge. The feasibility of this approach to automatic indexing was validated in a direct comparison with human indexing performance.

We have identified several general problem areas for further study. Among these are three primary issues: (1) the problem of domain knowledge and of integrating different domains with each other and with generic seed knowledge; (2) use of context for pronoun resolution, word disambiguation, and identification of knowledge domains; and (3) handling multiple text references underlying a single semantic relation from multidocument, multidomain databases and especially, deciding the priority of such references.

We also recommend that a knowledge-based interactive aid be designed and developed to assist an author in preparing the Research and Technology Work Unit Summary (form DD-1498).

LITERATURE CITED

(1) Landauer, T.K., Dumais, S.T., Gomez, L.M., and Furnas, G.W. "Human factors in data access." The Bell System Technical Journal, Vol. 61, No. 9, pp. 2487-2509, 1982.

(2) Griffin, A.F. "Intelligent information retrieval from on-line technical documentation." Proceedings of Joint Services Workshop on Maintenance Applications of Artificial Intelligence, Boulder, CO, Oct 4-6, 1983.

(3) Blair, D.C., and Maron, M.E. "An evaluation of retrieval effectiveness for a full-text document-retrieval system." Communications of the ACM, Vol. 28, No. 3, pp. 289-299, 1985.

(4) Salton, G., Buckley, C., and Fox, E.A. "Automatic query formulations in information retrieval." Journal of the American Society for Information Science, Vol. 34, No. 4, pp. 262-280, 1983.

(5) Tou, F.N., Williams, M.D., Fikes, R.E., Henderson, A., and Malone, T.W. "RABBIT: An intelligent database assistant." Proceedings of the National Conference on Artificial Intelligence. Pittsburg, PA, August 1982.

(6) Wang, Y., Vandendorpe, J., and Evens, M. "Relational thesauri in information retrieval." Journal of the American Society for Information Science, Vol. 36, No. 1, pp. 15-27, 1985.

(7) Eisenberg, M. The direct use of online bibliographic information systems by untrained end users: A review of research. Information Resources Publications, Syracuse, N.Y., 1983.

(8) Quillian, M.R. "Semantic memory." In M. Minsky (Ed.), Semantic Information Processing. MIT Press, Cambridge, Mass., 1968.

(9) Brachman, R.J. "On the epistemological status of semantic networks." In N.V. Findler (Ed.), Associative Networks--Representation and Use of Knowledge by Computers. Academic Press, N.Y., 1979.

(10) Kiersey, D.M. Word learning with hierarchy-guided inference. Ph.D. Dissertation, Department of Computer Science, University of California, Irvine, 1983.

(11) Fahlman, S.E. "Representing and using real-world knowledge." In P. Winston & R.H. Brown (Eds.), Artificial Intelligence: An MIT Perspective, pp. 454-470. MIT Press, Cambridge, Mass, 1979.

(12) Jones, K.P. "How do we index: A report of some ASLIB informatics group activity." The Journal of Documentation, Vol. 39, No. 1, pp. 1-23, 1983.

(13) Kucera, H. and Francis, N. Computational Analysis of Present-Day American English. Brown University Press, 1967.

(14) Dillon, M. and Gray, A.S. "FASIT: A fully automatic syntactically based indexing system." Journal of the American Society for Information Science, Vol. 34, No. 2, pp. 99-108, 1983.

(15) Fillmore, C. "The case for case." In E. Bach and R. Harms (Eds.), Universals in Linguistic Theory, pp. 1-88. Holt, Rinehart and Winston, New York, 1968.

(16) Charniak, E. "The case-slot identity theory." Cognitive Science, Vol. 5, No. 3, pp. 285-292, 1981.

(17) Schank, R.C. Conceptual Information Processing. North Holland Publishing, Amsterdam, 1975.

(18) Britton, B.K. "Lexical ambiguity of words used in English text." Behavior Research Methods & Instrumentation, Vol. 10, No. 1, pp. 1-7, 1978.

(19) Alshawi, H. A mechanism for the accumulation and application of context in text processing. Technical Report No. 48, University of Cambridge Computer Laboratory, 1983.

(20) Charniak, E. "Passing markers: A theory of contextual influence in language comprehension." Cognitive Science, Vol. 7, pp. 171-190, 1983.

(21) Kieras, D.E. "The potential for advanced computerized aids for comprehensible writing of technical documents." Technical Report No. 17 (TR85/ONR17), Office of Naval Research, 1985. (ADA150501)

DISTRIBUTION

ADMINISTRATOR
DEFENSE TECHNICAL INFORMATION CENTER
ATTN DTIC-DDA (12 copies)
CAMERON STATION, BUILDING 5
ALEXANDRIA, VA 22304-6145

US ARMY LABORATORY COMMAND
ATTN COMMANDER, AMSLC-CG
ATTN TECHNICAL DIRECTOR, AMSLC-CT
ATTN PUBLIC AFFAIRS OFFICE, AMSLC-PA
2800 POWDER MILL ROAD
ADELPHI, MD 20783-1145

INSTALLATION SUPPORT ACTIVITY
ATTN RECORD COPY, SLCIS-IM-TS
ATTN LIBRARY, SLCIS-IM-TL (3 copies)
ATTN LIBRARY, SLCIS-IM-TL (WOODBRIDGE)
ATTN TECHNICAL REPORTS BRANCH, SLCIS-IM-TR
ATTN LEGAL OFFICE, SLCIS-CC
2800 POWDER MILL ROAD
ADELPHI, MD 20783-1145

HARRY DIAMOND LABORATORIES
ATTN D/DIVISION DIRECTORS
ATTN DIVISION DIRECTOR, SLCHD-RT
2800 POWDER MILL ROAD
ADELPHI, MD 20783-1197

END

FILMED

5-86

DTIC